Análise de Termos dos Títulos Publicados nos Anais do XXI ENANCIB por meio do *Software* Nvivo

Analysis of the terms of the titles published in the proceedings of the XXI ENANCIB using the NVivo software

Marcos de Souza (1)

(1) Universidade Federal de Minas Gerais, Brasil, marcosdesouza82@gmail.com.

Resumo

O objetivo geral da pesquisa foi analisar a eficiência na extração de termos por meio do *software* NVivo. Dentre os objetivos específicos, buscou-se identificar os termos mais frequentes contidos nos títulos dos Grupos de Trabalho (GTs), bem como compará-los aos extraídos do GT7 - Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação. Outro objetivo foi analisar *cluster* por similaridade entre os termos contidos nos títulos. Para o tratamento dos dados empíricos, foram realizadas as etapas: a) pré-análise, seleção e preparação do material; construção de *corpora*; b) exploração do material – técnicas de codificação; extração de termos; e c) tratamento dos resultados – operações estatísticas; interpretação, descrição e análise. Os termos com maior frequência em todos os GTs foram "informação, "análise", "ciência", "conhecimento" e "gestão". Já os termos do GT7 com maior frequência foram "ciência", "análise", "informação", "produção" e "coautoria". O termo de maior impacto do GT7 foi "produção", que representou 46,1% do total de frequência. A maior similaridade entre os termos ocorreu com os títulos dos GTs: 5 e 8; 2 e 7 e; 3 e 6. A extração de termos utilizando o *software* NVivo são ineficazes para análises profundas, uma vez que os resultados são apresentados no formato de unigramas e podem ficar fora de contexto quando analisados individualmente.

Palavras-chave: Extração automática de termos; Frequência de termos; Similaridade de *cluster*; Comunicação Científica

Abstract

The general objective of the research was to analyze the efficiency in the extraction of terms through the NVivo software. Among the specific objectives, we sought to identify the most frequent terms contained in the titles of the Working Groups (WGs), as well as to compare them to those extracted from WG7 - Production and Communication of Information in Science, Technology & Innovation. Another objective was to analyze cluster by similarity between the terms contained in the titles. For the treatment of empirical data, the following steps were carried out: a) pre-analysis, selection and preparation of the material; corpora construction; b) exploration of the material – coding techniques; term extraction; and c) treatment of results

– statistical operations; interpretation, description and analysis. The most frequent terms in all GTs were “information, “analysis”, “science”, “knowledge” and “management”. The terms of GT7 most frequently were “science”, “analysis”, “information”, “production” and “co-authorship”. The term with the greatest impact in GT7 was “production”, which represented 46.1% of the total attendance. The greatest similarity between the terms occurred with the titles of the GTs: 5 and 8; 2 and 7 and; 3 and 6. Extraction of terms using the NVivo software is ineffective for in-depth analyses, as the results are presented in unigram format and may be out of context when analyzed individually.

Keywords: Automatic term extraction; Term frequency; Cluster similarity; Scientific Communication

1 Introdução

A análise de um conjunto de documentos, denominado *corpus* – ou *corpora* quando existe mais de um conjunto de documentos –, pode ser realizada com base em diferentes métodos e técnicas que possibilitam, por exemplo, *insights* ou a identificação de perspectivas futuras sobre uma determinada área de conhecimento. A extração de termos de um conjunto de documentos é uma dessas técnicas ao identificar as palavras e suas respectivas frequências – contagem de quantas vezes um determinado termo apareceu em um conjunto de documentos.

Esse tipo de análise, antes da era tecnológica, era realizado de maneira manual e, dependendo do quantitativo de material a ser analisado, poderia levar meses ou anos. Com o advento as tecnologias computacionais na primeira metade do século XX e da computação eletrônica na década de 1970, *software* para análise de dados qualitativos e quantitativos aplicados em pesquisas, não só em Ciências Sociais Aplicadas, tornaram-se mais dinâmicos e precisos no tratamento dos dados e, consequentemente, em seus resultados.

Partindo desse princípio, pergunta-se: de que maneira os termos extraídos dos títulos das pesquisas científicas têm se apresentado nos anais do XXI Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ENANCIB)? Acredita-se como hipótese que a extração de termos por meio de combinações sequenciais de palavras pode apresentar resultados que contribuem para interpretações de conteúdos de maneira mais assertivas quando se comparado a extração de termos únicos, listados por meio de frequências.

Diante disso, o objetivo geral da pesquisa está em analisar a eficiência na extração de termos em *corpora* por meio do *software* NVivo. Dentre os objetivos específicos, estão: a) identificar os termos mais frequentes contidos nos títulos do Grupos de Trabalho (GTs) do

ENANCIB; b) comparar a representação dos termos mais frequentes dos títulos do GT7 com a totalidade dos termos encontrados nos títulos dos demais GTs do ENANCIB; e c) analisar *cluster* por similaridade de termos contidos nos títulos dos GTs do ENANCIB, com ênfase no GT7. Faz necessário ressaltar que não faz parte dos objetivos da pesquisa identificar tendências temáticas dos *corpora* analisados.

Esta pesquisa justifica-se, uma vez que a análise da extração de termos contidos em um *corpus* ou *corpora* pode apresentar tendência sobre o desenvolvimento atual e/ou sobre perspectivas futuras para uma determinada área de conhecimento. A escolha do GT7 do ENANCIB como paramtro referencial de comparação com os títulos dos demais GTs, ocorre pelo alinhamento de seu ementário com os elementos constituintes no referencial teórico da pesquisa em questão.

Na primeira subseção do referencial teórico deste estudo, são apresentados os conceitos teóricos sobre pesquisa, linguagem científica, comunicação e divulgação científica. Na subseção seguinte, são apresentados os conceitos empíricos da pesquisa, como linguística de *corpus*, extração de termos, tokenização, n-gramas e análise de *cluster*.

2 Referencial teórico

2.1 Conceitos teóricos que norteiam a pesquisa

Enquanto atividade básica da ciência, a pesquisa é a descoberta científica da realidade, sendo a própria geração do conhecimento que antecede a transmissão do conhecimento. Como existem diferentes possibilidades para explicar determinada realidade, as formas humanas não esgotam a verdade. Assim, sempre há o que descobrir na realidade (Michel 2015).

A pesquisa é definida como um procedimento racional e sistemático, cujo objetivo está em buscar soluções para problemas propostos. A pesquisa pode surgir em dois momentos: a) quando não se dispõe de informações necessárias para solucionar um problema; e b) quando as informações disponíveis se encontram em desordem e não podem ser adequadas para solucionar o problema (Gil 2010).

O desenvolvimento da pesquisa ocorre ao longo de um processo constituído por diferentes fases, que vão desde a adequação do problema até a apresentação satisfatória dos resultados, utilizando-se de métodos e técnicas de investigação científica (Gil 2010).

A linguagem utilizada na pesquisa científica deve ser objetiva (Aquino 2010). Trata-se de uma linguagem técnica, com agrupamento de ideias sequências lógicas, cuja finalidade é transmitir o conhecimento. Além disso, deve ter uma linguagem coerente com as regras gramaticais, ser a mais didática possível, utilizar a impessoalidade no texto e evitar vocabulários populares, vulgares e pomposos (Marconi e Lakatos 2003). Outras características estão relacionadas como a organização e apresentação final do texto (Luiz 2018).

Refere-se à comunicação científica um conjunto de atividades associadas à produção, disseminação e uso da informação (Garvey e Griffith 1979). A comunicação científica é constituída por pesquisas científicas submetidas e aprovadas em eventos acadêmicos, como congressos, simpósios, fóruns, colóquios, encontros e reuniões. Esse tipo de pesquisa possui estrutura resumida, que pode variar de acordo com a instituição organizadora (Michel 2015). Outras formas que constituem a comunicação científica são os livros e capítulos de livros, artigos e resumos publicados em periódicos científicos (Mueller 2007).

Durante a construção da pesquisa, - que perpassar por diferentes etapas como escolha do tema, revisão de literatura, formualação do problema, determinação dos objetivos, construção da justificativa, elaboração de hipóteses, escolha da metodologia e descrição dos métodos, coleta, tabulação e análise de dados, análise e discussão dos resultados, considerações finais, redação e apresentação - é realizada a geração da informação, que, posteriormente, submetida aos crivos realizados em avaliação por pares e se aprovada, é disseminada conferindo a transparência da pesquisa por meio dos canais de comunicação, podendo ser formais ou informais, escritos ou orais (Garvey e Griffith 1979).

Dentro do contexto de pesquisas científicas submetidas e aprovadas em eventos científicos, fazem parte da estrutura de elementos o nome, local, data, patrocinador do evento, título, nome e credenciais do autor, resumo, conteúdo – introdução, desenvolvimento, conclusão – e referências. As pesquisas são expostas oralmente ou em forma de painéis e buscam a divulgação dos resultados – e não a profundidade da análise, considerando sua dinâmica (Michel 2015).

Já a divulgação científica se refere à "[...] utilização de recursos, técnicas, processos e produtos (veículos ou canais) para a veiculação de informações científicas, tecnológicas ou associadas a inovações ao público leigo" (Bueno 2009 p. 162). Constituem a divulgação científica: a) tornar a ciência mais compreensível; b) aproximar a sociedade da comunidade científica; c) levar conhecimento e ampliar o debate; e d) estimular o pensamento crítico (Albagli 1996).

Valeiro e Pinheiro (2008) destacam que a comunicação científica é responsável por estabelecer o diálogo entre o pesquisador e o público da comunidade científica. Já a divulgação científica busca a comunicação com a comunidade em geral.

O ENANCIB é considerado o principal evento brasileiro de pesquisa e pós-graduação na área da Ciência da Informação que busca discutir, refletir e compartilhar informações acerca da produção do conhecimento científico (ENANCIB, 2021a). O Quadro 1 apresenta os GTs e ementários do ENANCIB.

Quadro 1 – GTs e ementários do ENANCIB

Grupos de Trabalho	**Ementários**
GT1 – Estudos Históricos e Epistemológicos da Ciência da Informação	Estudos históricos e epistemológicos da Ciência da Informação (escolas de pensamento, correntes teóricas, autores e obras de fundamentação, leituras teórico-metodológicas e conceituações). Constituição, desenvolvimento e inovação conceitual, teórica e metodológica do campo científico informacional. Os objetos de estudos da Ciência da Informação e suas transformações teórico-conceituais. Reflexões e discussões sobre disciplinaridade, interdisciplinaridade e transdisciplinaridade.
GT2 – Organização e Representação do Conhecimento	Teorias, metodologias, políticas, instrumentos, processos e produtos para a organização e representação do conhecimento recuperação e acesso à informação, nas suas dimensões epistemológicas, aplicadas, sociais, culturais e terminológicas enquanto conhecimento socializado, institucionalizado ou não, em ambientes informacionais (tais como: arquivos, museus, bibliotecas e congêneres), incluindo o uso e desenvolvimento das tecnologias de informação e as relações inter, multi e transdisciplinares neles verificadas.
GT3 – Mediação, Circulação e Apropriação da Informação	Estudo dos processos e das relações entre mediação, circulação e apropriação de informações, em diferentes contextos e tempos históricos, considerados em sua complexidade, dinamismo e abrangência, bem como relacionados à construção e ao avanço do campo científico da Ciência da Informação, compreendido em dimensões inter e transdisciplinares, envolvendo múltiplos saberes e temáticas, bem com contribuições teórico-metodológicas diversificadas em sua constituição.

GT4 – Gestão da Informação e do Conhecimento	Gestão de ambientes, sistemas, unidades, serviços, produtos de informação e recursos informacionais. Estudos de fluxos, processos, usos e usuários da informação como instrumentos de gestão. Gestão do conhecimento e aprendizagem organizacional no contexto da Ciência da Informação. Marketing da informação, monitoramento ambiental e inteligência competitiva. Estudos de redes para a gestão. Aplicação das tecnologias de informação e comunicação à gestão da informação e do conhecimento.
GT5 – Política e Economia da Informação	Políticas e regimes de informação. Informação, Estado e governo. Propriedade intelectual. Acesso à informação. Economia política da informação e da comunicação. Produção colaborativa. Poder, ativismo e cidadania. Conhecimento, aprendizagem e inovação. Ética da informação. Informação e ecologia.
GT6 – Informação, Educação e Trabalho	O mundo do trabalho informacional: atores, cenários, competência em informação, dimensões e habilidades. Organização, processos de trabalho em dispositivos de informação e cultura. As relações entre informação, educação, trabalho, saúde e tecnologia. Regulamentação profissional, entidades sindicais, associações de classe e mercado de trabalho e competência profissional. Diversidade cultural, representações sociais, práticas e construção identitária dos profissionais da informação. Responsabilidade social, ética e profissional na Ciência da Informação. As bases curriculares e experiências pedagógicas: formação e perfil profissional ou docente.
GT7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação	Estudos teóricos, aplicados e metodológicos sobre a produção, comunicação e uso da informação em Ciência, Tecnologia e Inovação. Inclui pesquisas relacionadas aos processos de comunicação, divulgação, análise e formulação de indicadores para planejamento, avaliação e gestão em CT&I.
GT8 – Informação e Tecnologia	Estudos e pesquisas teórico-práticos sobre e para o desenvolvimento de tecnologias de informação e comunicação que envolvam os processos de geração, representação, armazenamento, recuperação, disseminação, uso, gestão, segurança e preservação da informação em ambientes digitais.
GT9 – Museu, Patrimônio e Informação	Análise das relações entre o museu (fenômeno cultural), o patrimônio (valor simbólico) e a informação (processo), sob múltiplas perspectivas teóricas e práticas de análise. Museu, patrimônio e informação: interações e representações. Patrimônio musealizado: aspectos informacionais e comunicacionais.
GT10 – Informação e Memória	Estudos sobre a relação entre os campos de conhecimento da Ciência da Informação e da Memória Social. Pesquisas transdisciplinares que envolvem conceitos, teorias e práticas do binômio 'informação e memória'. Memória coletiva, coleções e colecionismo, discurso e memória. Representações sociais e conhecimento. Articulação entre arte, cultura, tecnologia, informação e memória, através de seus referenciais, na contemporaneidade. Preservação e virtualização da memória social.
GT11 – Informação & Saúde	Estudos das teorias, métodos, estruturas e processos informacionais em diferentes contextos da saúde, considerada em sua abrangência e complexidade. Impacto da informação, tecnologias, e inovação em saúde. Informação nas organizações de saúde. Informação, saúde e sociedade. Políticas de informação em saúde. Formação e capacitação em informação em saúde.

Fonte: ENANCIB (2021b *online*)

No ano de 2022 foi incorporado ao ENANCIB o GT12 – Informação, Estudos Étnico-Raciais, Gênero e Diversidades, tendo como ementa:

> Estudos teóricos e aplicados em informação sobre Raça, Classe, Gênero, Sexualidades e Interseccionalidades. Teorias Críticas, Culturais, Racial, Feministas e Queer. Correntes teóricas, escolas de pensamento, bases metodológicas-conceituais e aplicações técnico-científicas dos estudos étnico-raciais, de gênero e de diversidade. Teorias, discursos, saberes, atividades científicas e profissionais em ambientes informacionais comunitários, populares e organizacionais. Relações sociais, de poder e resistências. Epistemicídio, violências e insurgências. Estudos Pós-Coloniais, Decoloniais e Anticoloniais. Estudos Críticos da Branquitude. Justiça Social, Informacional, Racial e de Gênero (ENANCIB 2022 *online*).

2.2 Conceitos empíricos que compõem a pesquisa

A Linguística de *Corpus* refere-se à coleta e exploração de conjuntos de dados linguísticos textuais, coletados a partir da utilização de critérios cujo propósito está em servir para uma pesquisa de uma determinada língua ou fazendo uso de uma variedade linguística. A exploração do conteúdo é realizada por meio de evidências empíricas extraídas do computador (Sardinha 2000).

De acordo com o dicionário Aurélio (2021), o termo *corpus* é corpo, conjunto de documentos sobre um assunto ou tema. São exemplos de *corpus* não computacionais o *Corpus* Helenístico definido por Alexandre, o Grande, na Grécia Antiga. Na Antiguidade e na Idade Média, foi produzido *corpora* de citações da Bíblia (Sardinha 2000).

Sardinha (2004) destaca diferentes tipos de aplicações que podem resultar em estudos baseados em *corpora* eletrônico, como por exemplo, a função de autorresumo, sintetizadores de voz, tradutores e digitadores – recursos disponíveis de usuários de um sistema operacional como o do Windows ou da Apple. O autor ainda enfatiza os programas para manuseamento de *corpora*, como concordanciadores, etiquetadores e extratores de frequência.

Grande parte dos *corpora* são constituídos por publicações eletrônicas textuais em diferentes formantos, como artigos científicos, resumos, notas, relatórios, correspondência eletrônica, dentre outros. Considera-se assim uma forma natural de armazenar informações (Han e Kamber 2006). Esse conjunto de documentos pode apresentar um quantitativo representativo de conhecimento a ser utilizado para diferentes finalidades (Shaw et al. 2001).

Com isso, o conceito mineração de texto tem como o principal objetivo encontrar termos de relevância em *corpus* ou *corpora* com grandes volumes de documentos, bem como estabelecer padrões e relacionamentos com base na extração de frequência de termos. Dentre as principais técnicas utilizadas para mineração de texto estão: a) Processamento de Linguagem Natiral – utiliza o compuador para melhorar o entendimento da linguagem natural por meio de ténicas para processar textos; b) Recuperação da Informação – métodos semânticos e estatísticos para processar de maneira automática o texto de documentos e com isso encontrar quais documentos possuem a resposta para o questionamento realizado; c) Extração de Informação – busca partes relevantes de um texto e extrai informações específicas (Machado et al. 2010).

Para que a extração automática de termos de um determinado *corpus* seja útil, faz-se necessário o processo de descoberta do conhecimento (Tan et al. 2005). O processo de extrair conhecimento de dados textuais exige técnicas específicas e complexas. Na estruturação do processamento, deve considerar as características da linguagem natural utilizada, a fim de manter o conhecimento expresso no texto (Hotho et al. 2005).

Uma das etapas realizadas para a extração de termos é a de pré-processamento. Responsável por tarefas como a realização da limpeza dos textos, remove as figuras, tabelas e *stop words* (em português, palavras de parada) que não possuem relevância para o contexto. Com a execução dessa etapa, é possível obter uma representação da coleção de textos em formato estruturado, preservando assim as características do *corpus* original (Feldman e Sanger 2007).

A remoção das *stop words* é fundamental para reduzir o tamanho do *corpus*, melhorar o desempenho de processamento de linguagem natural e apresentar resultados concisos, independente da finalidade do estudo ou modelo a ser utilizado (Souza 2020). Um dos modelos utilizados para análise de textos é o *Bag of Words* (em português, saco de palavras). Nele, cada termo contido em um determinado *corpus* remete a um atributo da base textual. Dessa maneira, uma coleção de documentos é representada por uma matriz de documentos-termos, em que cada linha da matriz é representada por um documento e cada coluna, por um termo. As células da matriz são preenchidas com a frequência dos termos de cada documento (Salton e Buckley 1988).

Para isso, é necessário realizar uma etapa chamada tokenização. Nela os textos são divididos em palavras, símbolos e outros elementos convertidos em uma sequência de palavras

separadas por meio do tratamento de pontuação e espaçamento. Quando a linguagem escrita é armazenada em um documento de computador, passa a ser representada por sequência de *strings* de caracteres (Souza 2020).

Nos campos da linguística computacional e da probabilidade, um n-grama é uma sequência contígua de n itens de uma determinada amostra coletada em *corpus* de texto ou de fala. Os itens podem ser fonemas, sílabas, letras, palavras ou pares de bases extraídos de acordo com a aplicação. Um n-grama é referido como unigrama, dois n-gramas como bigramas e três n-gramas como trigramas. A partir daí, refere-se como quatro grama, cinco grama, seis grama etc. (Broder et al. 1997). Esses itens podem ser utilizados em modelos como o *Bag of Words.*

Por fim, a análise de *cluster* diz respeito à classificação de objetos em diferentes grupos. Essa classificação deve apresentar alguma semelhança entre tais objetos, considerando as funções de distância estatística. Além disso, deve ocorrer de maneira automática e não supervisionada – sem considerar previamente as características do *corpus* e sem realizar testes que possam tendenciar a classificação (Jain et al. 1999; Theodoridis e Koutroumbas 1998).

As similaridades entre os grupos são determinadas de maneira a obter homogeneidade dentro dos grupos e heterogeneidade entre eles. Diferentes técnicas foram desenvolvidas para auxiliar a formação dos agrupamentos por similaridade, tendo em vista a dificuldade de examinar grandes volumes de dados (Doni 2004).

Zaiane et al. (2002) destaca que uma análise de *cluster* criteriosa exige métodos que apresentam diferentes características, dentre elas: a) lidar com dados com alta dimensionalidade; b) ser escalável com o número de dimensões e com a quantidade de elementos a serem agrupados; c) definir agrupamentos de diferentes tamanhos e formas; d) lidar com diferentes tipos de dados; e) eficiente para trabalhar com ruídos; e f) apresentar resultado consistente.

Os conceitos teóricos que norteiam esta pesquisa estão relacionados ao fazer da construção de conteúdos dos *corpora* analisados, ou seja, as características necessárias para produzir textos científicos. A intenção é abordar especificamente os títulos de pesquisas científicas dos tipos resumos expandidos e trabalhos completos publicados nos anais do XXI ENANCIB. O referencial empírico aborda os conceitos referentes aos *corpora*, processamento de linguagem natural e

extração automática de termos que buscam responder ao problema, comprovar a hipótese e atingir os objetivos da pesquisa.

3 Metodologia

A pesquisa se classifica como aplicada quanto a sua finalidade, quantitativa quanto à abordagem do problema, exploratória do ponto de vista dos objetivos, e bibliográfica / experimental quanto aos procedimentos técnicos (Gil 2010). O referencial teórico foi elaborado a partir de livros e artigos científicos encontrados em bases de dados como SciELO, Google Scholar e o Portal de Periódicos da CAPES.

Para a pesquisa empírica, foram realizadas as seguintes fases: 1) organizar em etapas – pré-análise, seleção e preparação do material; construção de *corpora*; 2) explorar o material – técnicas de codificação; extração de termos; e 3) tratamento dos resultados – operações estatísticas; síntese e inferências; interpretação e descrição.

A fim de construir os *corpora*, foram coletados os títulos dos artigos completos e resumos expandidos publicados nos anais do XXI ENANCIB, totalizando 11 documentos com os títulos dos respectivos GTs a serem analisados. O ENANCIB foi escolhido para compor os *corpora* mediante a sua representatividade de Pesquisa e de Pós-graduação da área de Ciência da Informação do Brasil.

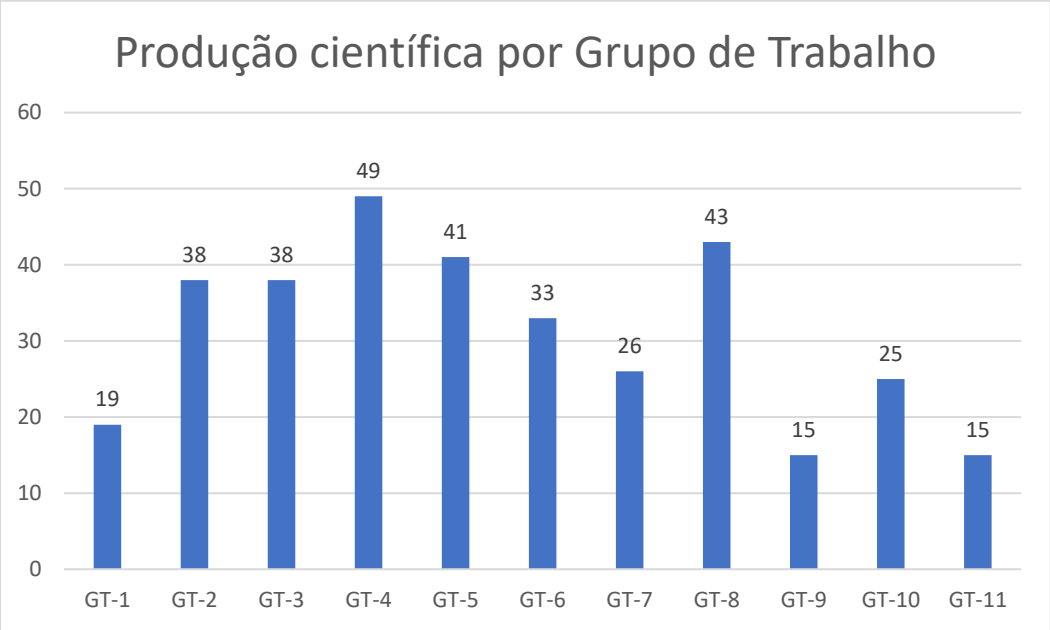
A limitação dos *corpora* se apresenta pelos seguintes fatores: a) quantidade representativa de títulos para serem analisados; b) trabalhos com menos de um ano de publicação, ou seja, com temas atuais e não sendo possível realizar uma análise diacrônica de comportamento de termos.

Os documentos foram organizados, explorados e tratados por meio do *software* NVivo. Todos os arquivos que constituem a parte empírica da pesquisa estão disponíveis para *download* no Figshare.

4 Resultados e discussões

Os anais do XXI ENANCIB totalizaram 342 trabalhos publicados nas modalidades de resumos expandidos e trabalhos completos. O GT1 apresentou 19 (5,6%) trabalhos do total das pesquisas; GT2, 38 (11,1%); GT3, 38 (11,1%); GT4, 49 (14,3%); GT5, 41 (12,0%); GT6, 33 (9,6%); GT7, 26 (7,6%); GT8, 43 (12,6%); GT9, 15 (4,4%); GT10, 25 (7,3%); e GT11, 15 (4,4%), conforme apresentado no Gráfico 1.

Gráfico 1 – Produção cientifica por GT do XXI ENANCIB



Fonte: Elaborado pelo autor (2022).

Os *corporas* constituídos com base nos títulos dos anais do XXI ENANCIB apresentam as seguintes características: 4.362 palavras; média de 12 palavras por título; maior título contendo 33 palavras; e menor título contendo 5 palavras. Já as características apresentadas pelo *corpus* com os títulos do GT7 são: total de 26 títulos de pesquisas publicadas; 341 palavras; média de 13,12 de palavras por título; maior título contendo 22 palavras; e menor título contendo 7 palavras.

Foram identificados um total de 1.289 termos únicos quando analisados todos os títulos publicados nos anais do XXI ENANCIB. Para esse quantitativo de palavras que compõem 342 títulos, foi utilizada, no *software* NVivo, a configuração de comprimento mínimo de duas letras

por palavra, argumento de correspondências exatas como “talk” e desconsiderados ajustes de argumentos como palavras derivadas (“talking”), sinônimos (“speak”), especialização (“whisper”) ou generalizações (“comunidade”).

Quando realizado o ajuste da análise dos *corpora* no NVivo para o cumprimento mínimo de quatro letras por palavra, eliminação das *stop words* e limite para exibir as 50 palavras mais frequentes, tornou-se possível identificar termos mais específicos, associados às temáticas do evento.

A Figura 1 apresenta nuvens de palavras dos títulos das pesquisas publicadas nos anais do XXI ENANCIB. A Figura 1-A representa os títulos com o tratamento mínimo e a Figura 1-B, os títulos com um maior refinamento no tratamento.

Figura 1 – Termos dos títulos dos anais do XXI ENANCIB

Fonte: Elaborado pelo autor (2022)

Dentre os termos mais frequentes que possuem um tratamento mínimo, representado na Figura 1-A, estão: “de”, com 329 repetições; “da”, com 256; “informação”, com 130; “em”, com 91; “na”, com 70; “para”, com 65; “uma”, com 46; “análise”, “ciência” e “dos”, com contagem de 40 para cada termo. Já os termos mais frequentes, que possuem um tratamento mais refinado, representado na Figura 1-B, estão: “informação”, com 70 repetições; “análise”, com 40; “ciência”,

com 40; “conhecimento”, com 26; “gestão”, com 24; “memória”, com 24; “mediação”, com 21; “biblioteca”, com 20; “dados”, com 19; e “Brasil”, com 17.

A Figura 2 apresenta nuvens de palavras constituída a partir dos termos contidos no *corpus* com os títulos do GT7, sendo a Figura 2-A para os termos com um tratamento mínimo e a Figura 2-B com um tratamento mais refinado, seguindo as mesmas configurações quando analisado os *corpora* contendo todos os títulos dos GTs.

Figura 2 – Termos dos títulos do GT7 do XXI ENANCIB

2-A **2-B**

Fonte: Elaborado pelo autor (2022)

Dentre os termos mais frequentes do GT7 com o tratamento mínimo, estão: “da” e “de”, com frequência de 26 cada; “em”, com 11; “na”, com 9; “ciência”, com 7; “análise”, “das”, “informação” e “produção”, com 6 cada; e “coautoria”, com 4, conforme representado na Figura 2-B. Já os termos mais frequentes, com tratamento mais refinado, conforme representado na Figura 2-B, estão: “ciência”, com 7 repetições; “análise”, “informação” e “produção”, com 6 cada; “coautoria”, com 4 repetições; e “aberto”, “acesso”, “pesquisa” e “publicações”, com 3 cada.

Quando analisados os resultados das frequências de termos submetidos ao tratamento mais refinado, foi possível identificar que essas palavras possuem características da área estudada,

eliminando, assim, os artigos, pronomes, as preposições, conjunções, dentre outras possibilidades, como as palavras que estejam fora do contexto.

O termo “ciência” apresentou o maior número de repetições na lista de termos refinados do GT7. Foi o terceiro termo mais frequente na lista de todos os grupos. O termo identificado no GT7 representa 17,5% do total de 130 frequências encontradas em todos os GTs. Seguindo a mesma lógica, o segundo termo mais frequente no GT7 foi “análise”, que também ocupa a mesma posição na lista de frequência de todos os grupos de trabalhos. Dessa maneira, o termo “análise”, no GT7, representa 15% do total de 40 repetições encontradas em todos os GTs.

O termo “informação” foi o terceiro mais frequente encontrado no GT7, sendo o mais frequente de todo o *corpus*. O termo equivale a 4,6% de um total de 130 repetições. O quarto termo mais frequente do GT7 foi “produção”, que ocupa a 18ª posição na lista com os termos de todos os GTs. O termo no GT7 representa 46,1% do total de 13 frequências em todo o *corpus*. O quinto termo mais frequente do GT7 foi “coautoria”, que não aparece entre os 50 termos mais frequentes de todo o *corpus*.

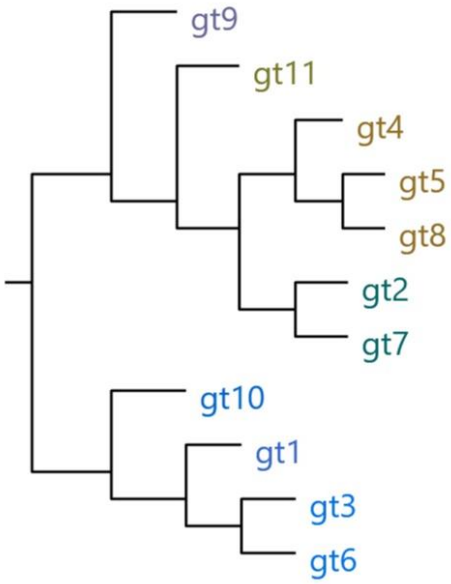
Os resultados contendo os termos que passaram pelo tratamento com maior refinamento, como a eliminação das *stop words*, apresentam maior representatividade quando se comparado aos termos que passaram por um tratamento mínimo de dados. Esses, por sua vez, carregam um excesso de palavras que não contribuem para a análise.

Os termos com maior frequência extraídos dos títulos das pesquisas científicas do GT7 foram “ciência”, “análise”, “informação”, “produção”, “coautoria”, “aberto”, “acesso”, “pesquisa”, “pesquisa” e “publicações”. Tais termos estão alinhados à proposta apresentada no ementário do GT7, pois abordam assuntos específicos da temática ou indicam perspectivas futuras para o GT. Entretanto, faz-se necessário ressaltar que os termos extraídos são genéricos, dificultando a interpretação dos dados, mesmo que seja por um especialista da área. Por exemplo, o termo “informação”, representado por um unigrama, pode abordar diferentes assuntos, como “Ciência ~~da~~ Informação”, “Gestão ~~da~~ Informação” ou “Informação ~~e~~ Sociedade” para bigramas ou “Gestão ~~da~~ Informação ~~e~~ ~~do~~ Conhecimento”, “Informação, Sociedade ~~e~~ Conhecimento” ou “Produção ~~e~~ Comunicação ~~da~~ Informação” para trigramas.

A solução para uma análise mais aprofundada dos termos contidos nos títulos do GT7 é a extração de termos realizada também por bigramas ou trigramas, como “Comunicação Científica” ou “Produção ~~e~~ Comunicação ~~da~~ Informação”, que apresentam assuntos ou áreas específicas. Contudo, essa não é uma solução disponibilizada pelo *software* NVivo.

Com relação à análise de *cluster* por similaridade de palavras contidas nos títulos das pesquisas, foi possível identificar uma proximidade entre os GTs 5 e 8 (amarelo), 2 e 7 (verde), 3 e 6 (azul) conforme apresentado na Figura 3. Cabe ressaltar que a análise de *cluster* é uma técnica exploratória que possibilita a visualização de padrões semânticos que agrupam fontes ou nós constituídos por palavras, atributos ou valores semelhantes. Dessa maneira, grupos próximos de fontes ou nós apresentam maior similaridade quando comparado-os com grupos que possuem maior distanciamento.

Figura 3 – Análise de *Cluster* por similaridade dos anais do XXI ENANCIB



Fonte: Elaborado pelo autor (2022)

Essa correlação é representada por meio de um dendrograma (ou diagrama de árvore) e exibe os grupos formados por agrupamento de observações em cada passo e em seus níveis de similaridade. A similaridade se difere por cores e colchetes. Quanto maior a distância, menor a similaridade e, quanto menor a distância, maior a similaridade.

Também é possível observar dois grandes grupos de similaridade no dendograma, sendo o primeiro constituído pelos GTs 9, 11, 4, 5, 8, 2 e 7 (ordem *top down* apresentada na Figura 3), que apresentam um maior distanciamento entre os colchetes e uma variedade de cores – roxo, amarelo e verde, caracterizando assim uma menor similaridade entre os títulos. Já o segundo grupo, constituído pelos GTs 10, 1, 3 e 6, apresentam um menor distanciamento e uma única cor – azul, caracterizando assim uma maior similaridade entre os títulos analisados.

A aproximação entre os GTs está relacionada ao conteúdo analisado, especificamente, aos títulos dos artigos completos e resumos expandidos publicados nos anais do XXI ENANCIB. Torna-se importante ressaltar que, tanto a similaridade de *cluster* quanto a extração automática de termos dizem respeito aos conteúdos dos *corpora*. Dessa maneira, diferente dos títulos, se fosse analisado somente os resumos das comunicações científicas ou mesmo as pesquisas na íntegra publicadas nos anais do evento, os resultados seriam diferentes por conta do conteúdo.

5 Considerações finais

Apesar de o NVivo ser um *software* para análises qualitativas, possui recursos específicos para análises quantitativas que apresentam certas limitações, especificamente ao abordar a extração automática de termos e suas respectivas frequências que possibilitam diferentes tipos de análises, dentre elas, os assuntos específicos abordados em um *corpus* ou mesmo a possibilidade de identificar diferentes tipos de *insights*, como por exemplo, perspectivas futuras de uma determinada área do conhecimento.

Quanto ao problema, os termos extraídos dos títulos das pesquisas científicas publicadas nos anais do XXI ENANCIB, por meio do *software* NVivo, apresentam características genéricas, dificultando as análises profundas, uma vez que os resultados são apresentados no formato de unigramas e podem ficar fora de contexto quando analisados individualmente.

Quanto ao objetivo geral e dentre a proposta do *software* NVivo, pode-se considerar eficiente a extração de frequência de termos dos *corpora*, ou seja, faz o que se propõe. Entretanto, quando se fala em análise de dados, os resultados podem ser rasos e ineficazes dependendo do objetivo da pesquisa, uma vez que a extração automática de frequência ocorre somente no formato

de unigrama. A implementação de opções para extração de frequência de termos por meio de bigramas e trigramas é uma solução que amplia a possibilidade de novas pesquisas com resultados e análises mais assertivas, uma vez que os termos nos formatos de bigramas e trigramas apresentam características mais específicas e menos generalistas quando comparado-os aos unigramas.

Os objetivos da pesquisa foram atingidos, dentre eles, destacam-se os percentuais representativos encontrados nos principais termos do GT7 quando comparado aos demais GTs. Dessa maneira, o termo com maior relevância no GT7 foi "produção", que representa 46,1% quando comparado à frequência do mesmo termo nos demais GTs. Já o termo com maior frequência extraído do GT7 foi "ciência", que representa 17,5% quando comparado à frequência dos demais GTs.

Quanto à análise de *cluster*, foi possível identificar uma maior similaridade de termos entre: a) GT5 – Política e Economia da Informação e GT8 – Informação e Tecnologia; b) GT2 – Organização e Representação do Conhecimento e o GT7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; c) GT3 – Mediação, Circulação e Apropriação da Informação e GT6 – Informação, Educação e Trabalho.

Essa proximidade de termos pode ocorrer tanto por termos genéricos que podem fazer parte do contexto de quaisquer grupos, quanto para termos específicos, comuns em ambos os GTs, não fazendo parte desse contexto as *stop words*.

A hipótese da pesquisa foi comprovada, embora a extração de termos do tipo bigrama ou trigrama não tenha sido realizada, já que o *software* NVivo não possui tais funcionalidades. Entretanto, fica evidente que termos únicos, como "Produção", mesmo passando por refinamento como a eliminação das *stop words*, apresentam características generalistas, dificultado, assim, a análise dos resultados. Isso seria diferente em uma extração de termos dos tipos bigramas e ou trigramas, como "Produção Cultural" ou "Produção ~~e~~ Comunicação ~~da~~ Informação", que se refere a uma sequência de termos específica e que delimita o assunto abordado.

Dessa maneira, sugere-se para pesquisas futuras sobre extração de termos em *corpora* ou *corpus*, a buscar por alternativas em outros *software* ou que reutilizem códigos-fonte de linguagem de programação disponibilizados na comunidade acadêmica, como por exemplo, na pesquisa de

Souza (2020), que disponibilizou um código fonte desenvolvido na linguagem de programação Python para extração de unigramas, bigramas, trigramas e suas respectivas frequências de termos em conjunto de documentos.

Importa ressaltar que limitações de conteúdo que constituem um *corpus* podem interferir na análise e interpretação dos resultados, como identificar *insights* ou tendências da área quando realizada a extração de frequência de palavras. Entende-se por limitações de conteúdo os *corpora* ou *corpus* com informações insuficientes para realizar diferentes tipos de análises.

Dessa maneira, sugere-se também a realização de novas pesquisas dentro do mesmo contexto, entretanto, com *corpora* de documentos contendo maior número de informações, como por exemplo a extração automática de termos de resumos ou dos próprios trabalhos acadêmicos - na íntegra - publicados nos anais do ENANCIB, bem como uma análise diacrónica de comportamento dos termos extraídos ao longo dos anos.

Embora os termos extraídos dos títulos publicados nos anais do XXI ENANCIB apresentem características genéricas, pode ocorrer também a existência de unigramas com maior especificidade, como por exemplo, "bibliometria", "*fakenews*" ou "COVID", que contribuem para uma análise e interpretação dos dados mais assertiva.

Notas

(1) NVivo: Software de análise de dados qualitativos produzido pela QSR International.

(2) Dados da pesquisa. Disponível em: https://doi.org/10.6084/m9.figshare.c.5935864.v1/. Acessado 7 jul. 2022

(3) Código fonte para extração de termos. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/. Acessado 7 jul. 2022.

Referências

Aquino, Italo de Souza. *Como escrever artigos científicos: sem "arrodeio" e sem medo da ABNT*. São Paulo: Saraiva, 2010.

Albagli, Sarita. "Divulgação científica: informação científica para cidadania". *Ciência da informação*, v. 25, n. 3, 1996, pp. 396-404, http://revista.ibict.br/ciinf/article/view/639/. Acessado 11 jul. 2022.

Broder, Andrei Z., et al. “Syntactic clustering of the web”. *Computer networks and ISDN systems*, [S.l.], v. 29, n. 8-13, 1997, pp. 1157-1166. https://www.sciencedirect.com/science/article/abs/pii/S0169755297000317/. Acessado 10 jul. 2022.

Bueno, Wilson da Costa. "Jornalismo científico: revisitando o conceito". *Jornalismo científico e desenvolvimento sustentável.* São Paulo: All Print 2009, pp. 157-178.

"Corpus". *DICIO, Dicionário Online de Português*, 7Graus, 2021, www.dicio.com.br/corpus. Acesso em: 11 jul. 2022.

Doni, Marcelo Viana. *Análise de cluster*: métodos hierárquicos e de particionamento, 2004. Universidade Presbiteriana Mackenzie, Trabalho de Conclusão de Curso, http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF/. Acessado 11 jul. 2022.

"ENANCIB". *Sobre - XXI ENANCIB – Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação*, 2021a, http://enancib2021rio.ibict.br/o-evento/sobre/. Acessado 10 jul. 2022.

"ENANCIB". *Grupos de Trabalho - XXI ENANCIB – Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação*, 2021b, https://enancib2021rio.ibict.br/o-evento/grupos-de-trabalho-gts/. Acessado 10 jul. 2022.

"ENANCIB". *GT 12 – Informação, Estudos Étnico-Raciais, Gênero e Diversidades*, 2022, https://www.ufrgs.br/enancib2022/programacao/gt-12/. Acessado 10 jul. 2022.

Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press, 2007.

Garvey, William D., and Belver C. Griffith. “Scientific communication as a social system”. *Communication: the essence of science*. Oxford: Pergamon, 1979. p.148-164. Appendix B.

Gil, Antonio Carlos. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 2010.

Han, Jiawei, and Micheline Kamber. "Data mining: concepts and techniques, 2nd". University of Illinois at Urbana Champaign: Morgan Kaufmann, 2006, http://hanj.cs.illinois.edu/bk2/bib/ch6bib.pdf/. Acessado 29 out. 2022.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. “A brief survey of text mining”. *Ldv Forum*, v. 20, n. 1, 2005, pp. 19-62, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.447.4161&rep=rep1&type=pdf/. Acessado 11 jul. 2022.

Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. “Data clustering: a review”. *ACM computing surveys (CSUR)*, v. 31, n. 3, 1999, pp. 264-323, https://dl.acm.org/doi/abs/10.1145/331499.331504/. Acessado 10 jul. 2022.

Luiz, Ercília Maria de Moura Garcia. *Escrita Acadêmica: princípios básicos*. Santa Maria: UFSM/NTE, 2018.

Machado, Aydano Pamponet., et al. "Mineração de texto em Redes Sociais aplicada à Educação a Distância". *Colabor@ - Revista Digital da CVA – Ricesu*, v. 6, n. 23, 2010, pp. 6-23, https://silo.tips/download/mineraao-de-texto-em-redes-sociais-aplicada-a-educaao-a-distancia/. Acessado 28 out. 2022.

Marconi, Marina de Andrade, and Eva Maria Lakatos. *Fundamentos da metodologia científica*. São Paulo: Atlas, 2013.

Michel, Maria Helena. *Metodologia e pesquisa científica em ciências sociais*: um guia prático para acompanhamento da disciplina e elaboração de trabalhos monográficos. São Paulo: Atlas, 2015.

Mueller, Suzana Pinheiro Machado. "A ciência, o sistema de comunicação científica e a literatura científica". Fontes de informação para pesquisadores e profissionais. Belo Horizonte: UFMG, 2007, pp. 21-34.

Salton, Gerard, and Christopher Buckley. “Term-weighting approaches in automatic text retrieval”. *Information processing & management*, v. 24, n. 5, 1988, pp. 513-523, https://www.sciencedirect.com/science/article/abs/pii/0306457388900210/. Acessado 11 jul. 2022.

Sardinha, Tony Berber. "Lingüística de corpus: histórico e problemática". *Delta: documentação de estudos em lingüística teórica e aplicada*, v. 16, n. 2, 2000, pp. 323-367, https://www.scielo.br/j/delta/a/vGknQkZQGsGYbrQfKmTZY4s/?format=pdf&lang=pt/. Acessado 10 jul. 2022.

Sardinha, Tony Berber. "Lingüística de Corpus: uma entrevista com Tony Berber Sardinha". *Revista Virtual de Estudos da Linguagem–ReVEL*, v. 2, n. 3, 2004, pp. 1-5, http://www.revel.inf.br/files/entrevistas/revel_3_entrevista_tony_berber_sardinha.pdf/. Acessado 10 jul. 2022.

Shaw, Michael J., et al. "Knowledge management and data mining for marketing". *Decision support systems*, v. 31, n. 1, 2001, pp. 127-137, https://www.sciencedirect.com/science/article/abs/pii/S0167923600001238/

Souza, Marcos de. *O comportamento de termos da Ciência da Informação por meio da modelagem de tópicos*, 2020, Universidade Federal de Minas Gerais, Tese, Acessado 28 out. 2022.

https://repositorio.ufmg.br/handle/1843/34292/. Acessado 10 jul. 2022.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. ”Introduction to Data Mining. ed”. *Addison-Wesley Longman Publishing Co., Inc.*, Boston, 2005.

Theodoridis, Sergios, and Konstantinos Koutroumbas. “System evaluation”. *Pattern recognition*. London: Academic Press, 1998. pp. 342-343.

Valeiro, Palmira Moriconi, and Lena Vania Ribeiro Pinheiro. “Da comunicação científica à divulgação”. *Transinformação*, v. 20, n. 2, 2008, p. 159-169, https://www.scielo.br/j/tinf/a/jXWgggxgBhXfsT57JDVbghp/abstract/?lang=pt#/. Acessado 10 jul. 2022.

Zaïane, Osmar R., et al. “On data clustering analysis: Scalability, constraints, and validation”. *Pacific-Asiad Conference on Knowledge Discovery and Data Mining. Springer*, Berlin, Heidelberg, 2002, pp. 28-39, https://link.springer.com/chapter/10.1007/3-540-47887-6_4/. Acessado 11 jul. 2022.

Received: 18/08/2022 Accepted: 20/12/2022